Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 58 (2015) 272 – 279

Procedia Computer Science

Second International Symposium on Computer Vision and the Internet(VisionNet'15)

Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition

Sharada V Chougule[a]*, Mahesh S Chavan[b]

[a] *Finolex Academy of Management & Technology, Ratnagiri, Maharashtra, India*
[b] *KIT's College of Engineering,Kolhapur, Maharashtra, India*

Abstract

The widespread use of automatic speaker recognition technology in real world applications demands for *robustness* against various realistic conditions. In this paper, a robust spectral feature set, called NDSF (Normalized Dynamic Spectral Features) is proposed for automatic speaker recognition in mismatch condition. Magnitude spectral subtraction is performed on spectral features for compensation against additive noise. A spectral domain modification is further performed using time-difference approach followed by Gaussianization Non-linearity. Histogram normalization is applied to these dynamic spectral features, to compensate the effect of channel mismatch and some non-linear effects introduced due to handset transducers. Feature extraction using proposed features is carried out for a text independent automatic speaker recognition (identification) system. The performance of proposed feature set is compared with conventional cepstral features like (mel-frequency cepstral coefficients and linear prediction cepstral coefficients), for acoustic mismatch condition caused by use of different sensors. Studies are performed on two databases: A multi-variability speaker recognition (MVSR) developed by IIT-Guwahati and Multi-speaker continuous (Hindi) speech database (By Department of Information Technology, Government of India). From experimental analysis, it is observed that, spectral domain dynamic features enhance the robustness by reducing additive noise and channel effects caused by sensor mismatch. The proposed NDSF features are found to be more robust than cepstral features for both datasets.

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)

* Corresponding author. Tel.: +91 9421142531
*E-mail address:*shardavchougule@gmail.com

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15)
doi:10.1016/j.procs.2015.08.021

*Keywords:*MFCCs (Mel Frequency Cepstral Coefficients), LPCCs (Linear Predictive Cepstral Coefficients, NDSF (Normalized Dynamic Spectral Features)

1. Introduction

Automatic Speaker Recognition is the task performed by the machine to recognize individual from a person's voice. The system can be text dependent (trained and tested for specific words or phrases) or text independent (no constrain on what is spoken). Depending on end task or decision it can be classified as speaker *identification* or *verification*. Ease of availability, naturalness and inexpensive devices (microphones) to collect the data are the main reasons of popularity of task specific applications related to speaker recognition. The potential applications of speaker identification can be found in multi-user systems such as *speaker tracking* or *speaker diarization* to locate the segment of a given speaker in an audio segment or in automatic segmentation of teleconferences. Also it founds useful in helping transcription of courtroom discussion and forensic applications.

The increasing use of speaker recognition system as a biometric and data-driven fields demands for accuracy in unconstrained situations. Robustness of speaker recognition system is one of the crucial issues for its use in real world applications. Any system is said to be *robust* if its performance varies negligibly due changes in uncontrolled conditions. Mismatch in training and testing conditions causes the original speech to be severely affected. This will drop the performance of the speaker recognition system and is always undesirable. In this view, robust front end features can enhance the accuracy of the system in any type of mismatch and noise.

Cepstral features are most commonly used in state-of-art speaker recognition. Ease of computation and satisfactory performance in clinical environment are the two main reasons of its widespread use. However studies [3, 4] found that the performance of speaker recognition system using these features is much degraded in actual working conditions, than tested in clinical environment. The reason for degradation is the sensitivity of cepstral features for noise and mismatch**.** Delta and double-delta features [5] are commonly appended to cepstral features to improve the recognition accuracy for speech and speaker recognition. A spectral level modification is proposed in [6], to improve speech recognition performance in the presence of noise and reverberation. Considering the similar approach, we propose modified spectral features called *Normalized Dynamic Spectral Features* (NDSF), to improve the robustness of speaker recognition in mismatched conditions. As the goal is to investigate the performance of the features alone, all the remaining system parameters are kept static in order to get unbiased results.

The paper is organized as follows: In section II a brief overview of various features and approaches towards robustness is reviewed. Computation of proposed NDSF features is described in section III. Description of database and system structure is described in IV with results and discussion in section V and conclusion given in Section VI.

2. Cepstral features and approaches towards robustness

The speaker specific information present in the speech signal can be observed in both physiological and behavioural characteristics. Accordingly, the speech features are categorized as *low level* and *high level* features [3]. Low level features (like spectral and cepstral features) contribute the information related to physical structure of vocal tract, while high level features (like prosodic, phonetic or conversational patterns) carry the behavioural characteristics of the speaker. State of art MFCC features using GMM speaker models were well studied in [7]. In [8], problem of speaker identification and verification in noisy conditions is studied by using missing-feature theory to model noise with unknown temporal-spectral characteristics. Combining LP residual phase information related to excitation source with cepstral features (MFCCs) representing vocal tract, had given improved speaker verification performance on NIST database [9]. In [10], use of simple prosodic features such as those extracted from fixed size contour segments, without the knowledge of any higher level information are fused with state-of-art cepstral based features showed improved performance for speaker recognition. Experimental evaluations of mismatched and limited data, with additive noise environment is carried out based on missing data approach using various channel compensation techniques like CMN and RASTA. Different modelling techniques like GMM, GMM-UBM and GMM-SVM are used in [11] to improve the robustness of text independent speaker identification.Cepstral based feature vector warping using a Gaussian target distribution is proposed in [12] to compensate for channel mismatch, additive noise and non-linear disturbances due to handset transducer. Authors in [13] investigated some of the newly proposed features like mel line spectral frequencies (MLSF), residual phase, mean and difference mean energy within the critical band (MECB) on NIST SRE 2001 database. A new set of feature vector set named mean Hilbert

envelope coefficients (MHEC) of Gamma-tone filter bank outputs is proposed in [14] to study speaker identification in reverberant mismatch condition. Improved channel robust MFCCs are proposed in [15] to compensate transducer/microphone mismatch for continuous speech speaker recognition. High energy regions are considered to be most reliable in the presence of noise. Front end processing based on autoregressive models followed by modulation filtering process is proposed for robust feature extraction in noisy conditions [16]. Recently, use of discrete Karhunen-Love transform (D-KLT) is made in order to reduce the dimensionality of conventional MFCCs [17], over short sequence frames for speaker identification.

3. NDSF Feature Analysis and Computation

Speech is a non-stationary (slowly time varying) signal when observed over a short interval of time (5 -100 ms) and its characteristics changes rapidly over the interval 100 ms -5 sec. In order to examine the speech signal , it is necessary to divide it into short frames of 20 to 30 ms. Analysis of speech signal done over this interval is known as *segmental* analysis. During this interval the characteristics of speech signal are assumed to be stationary. Short-time Fourier transform (STFT) is commonly used to explore the spectral contents of speech signal. A window function (usually hamming window) is used to avoid spectral leakage caused due to direct framing. The convenient short time stationary behavior is exploited to characterize vocal tract transfer function unique to a particular speaker. From source-filter model of speech production [18], rapidly changing excitation source provides basic temporal fine structure; while slowly varying filter provides spectral color (or spectral shape). The peaks of the vocal tract response correspond approximately to its *formants*, which depends on shape and length of vocal tract providing useful information related to a particular speaker.

The speech signal gets deteriorated (or distorted) while travelling through various transmission channels. Communication channels or recording devices may introduce noise in the original signal. Such noise is known as convolutive noise and is observed as additive in frequency domain. At the first stage, input speech signal is pre-emphasized using a simple first order FIR filter, to avoid the spectral tilt [18] due to nature of glottal pulses. Hamming window of 25 msec with a frame shift of 10 msec is used for framing the speech signal. Short-time Fourier transform is then applied on the windowed speech, to obtain the spectral details. The Fourier transform of the windowed speech segment $X(\omega,\tau)$ is given by:

$$X(\omega,\tau)=\frac{1}{P}\sum_{k=-\infty}^{\infty}H(\omega_k)G(\omega_k)W(\omega-\omega_k,\tau) \tag{1}$$

where $H(\omega_k)G(\omega_k)$ represents the spectral envelope consisting of glottal and vocal tract contribution and $W(\omega,\tau)$ is the Fourier transform of window function. The global shape of spectral envelope is considered to be most informative part of the spectrum in speaker recognition.

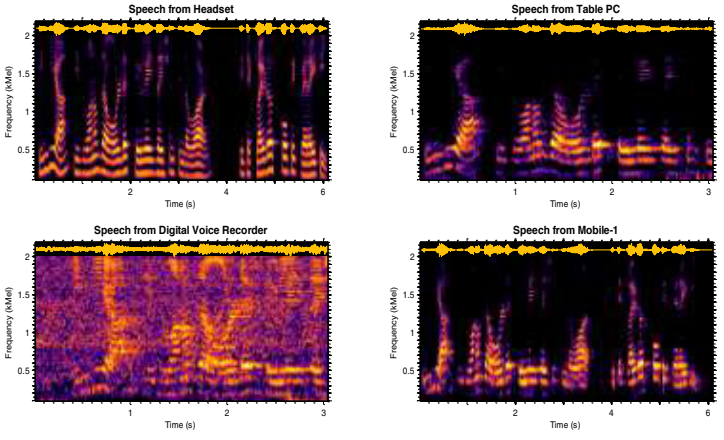


Fig.1 Spectrogram of windowed Speech signal (same speaker) from different sensors

Fig.1 shows the narrowband spectrogram of the same speaker's speech recorded with different sensors from IITG-MVSR database. It is observed that, the spectral details are clear and distiguishable for the speech collected from the Headset, whereas much distorted for the one collected from Digital Voice Recorder (DVR). Thus the type of device/handset may cause acoustic mismatch degrading the speech to be processed. Consequently, the speaker specific features will also get damaged, degrading the overall system performance.

To recuperate the speech signal convolved with the impulse response of communication channel or a recording devices, *spectral subtraction* (SS) is performed on the magnitude spectrum of the speech signal [19]. In this technique, the estimate of average magnitude spectrum is subtracted from the magnitude spectum of speech distorted through the channel. The spectral subtraction filter can be expressed as the product of noisy (distorted) speech spectrum $Y(k)$ and spectral gain function, $W_{SS}(k)$ as:

$$X'(k) = W_{ss}(k)Y(k) \tag{2}$$

where the frequency response of spectral subtraction filter $W_{SS}(k)$, is given as:

$$Wss(k) = fn\left[1 - \frac{\alpha(k)N'(k)}{Y(k)}\right] \tag{3}$$

In equation (3), $N'(k)$ represents the estimate of noise average amplitude spectrum, $\alpha(k)$ is the frequency dependent subtraction factor. The function $fn(.)$ can be chosen to avoid negative values of $W_{SS}(k)$ and provide smoother frequency response [20].

A set of overlapping band-pass filters (40 filters over the sampling frequency of 16 kHz) with *mel-scale* spacing [1] is used to do energy integration over the entire frequency band. These filters extract *subband energy* as a spectral feature from every speech frame. Since short time speech power changes faster than short-time noise power, the time difference (delta) operation performed on each frame can attenuate the noise components. Thus, the spectral domain dynamic information helps to boost the rapidly changing speech components and suppress slowly varying noise components. The delta parameter of *i*th feature is defined as:

$$\Delta f_k[i] = f_{k+M}[i] - f_{k-M}[i] \tag{4}$$

where $f_k[i]$ denote *i*th feature in the *k*th time frame with M typically 2-3 frames. We call these features as dynamic spectral features. The dynamic spectral features trim down any additive noise and help to improve robustness of the features.

Direct use of dynamic spectral features is unsuitable for speaker recognition application, due to their non-Gaussian nature. *Feature warping* is performed further, to map original spectral sample space into Gaussian distribution. The speaker dependant feature distribution is mapped to target distribution using histogram normalization, which is similar in concept to histogram equalization used in image analysis. Each spectral feature vector is warped to Gaussianized distribution (on frame by frame basis) so that the resulting feature distribution is made more consistent across the channel and sensor mismatch.

A normal distribution in a variable X with mean μ and variance is a statistic distribution with probability distribution function is given by [21]:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{5}$$

The source spectral features with measured distribution $f(y)$ are being mapped to warped features component having Gaussianized distribution $h(z)$ as:

$$h(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{6}$$

Fig.2 shows the histogram of original and warped spectral features (before and after Gaussianization) for a single speaker.

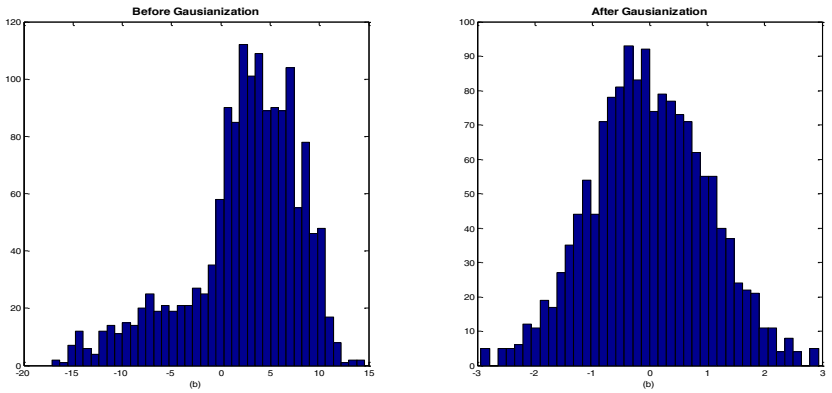


Fig.2 .Dynamic spectral features (a) Before and (b) After Gaussianization

The Gaussianized cepstral coefficients are normalized using cepstral mean normalization (CMN) to linear channel effects locally. This temporal mean is a rough estimate of the channel response. Acceleration coefficients are then derived from dynamic spectral features in cepstral domain, written from equation (4) as:

$$\Delta\Delta f_k[i] = \Delta f_{k+M}[i] - \Delta f_{k-M}[i] \tag{7}$$

Fig.3 shows the important steps in the extraction of the NDSF features discussed above.

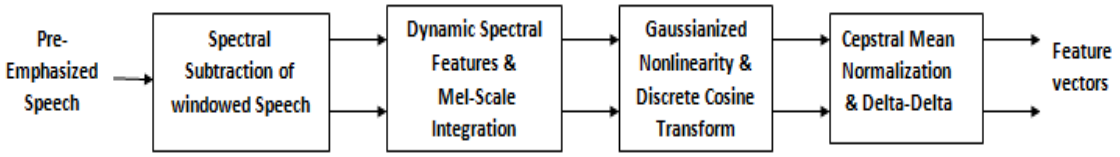


Fig.3. Block diagram showing the extraction of NDSF features

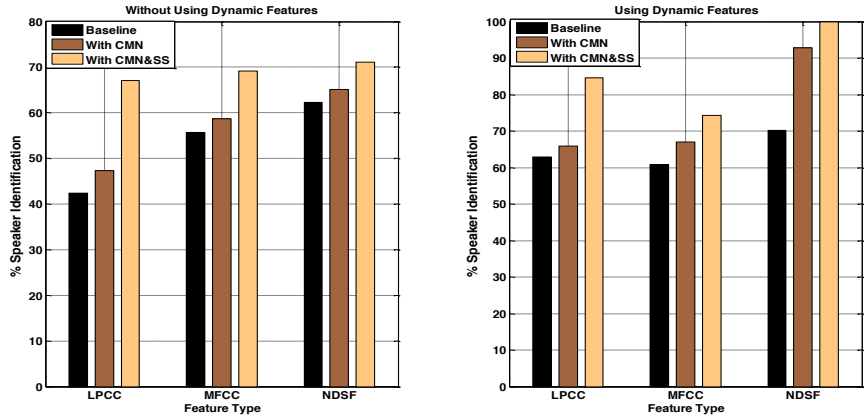
4. System Structure and Database

For experimental analysis a closed set text independent speaker identification system is build with LBG vector quantization as pattern formation technique [22]. The model formation is based on three separate features namely: LPCC, MFCC and NDSF respectively. Vector quantization is computationally efficient with storing and comparing large amounts of template data in the form of individual spectra. A multi-variability speaker recognition database with sensor mismatch from EMST Lab [23] is used for experimental analysis. Speech data from 100 speakers (81 Male and 19 Female) is collected using five different sensors namely Headset, Table PC, Digital voice recorder, Mobile-1 and Mobile-2. As speech from each device is sampled at different sampling rate, all the speech files are re-sampled at 8 kHz. The speech is of read style in English for 3-5 minutes duration. Out of which first 30 ms speech is

used for training the system and 10 ms speech from the same speech file is chosen arbitrarily for testing. The second database is multi-speaker, continuous (Hindi) speech database generated by TIFR, Mumbai (India) and made available by Department of Information Technology, Government of India. The database contains a total of approximately 1000 Hindi sentences, a set of 10 sentences read by each of 100 speakers. These 100 sets of sentences were designed such that each set is `phonetically rich' [24]. The speech data was simultaneously recorded using two microphones: one good quality, close-talking, directional microphone and another desk-mounted Omni-directional microphone. Database of 97 speakers is used for evaluation purpose. Training and testing is performed with different sensors (mismatch). The database used in this experimentation allows us to well investigate the mismatch effect on the performance of speaker recognition system. Out of various mismatch conditions available in the dataset, we have cosidered only the case of sensor mismatch.

5. Results and Discussion

From experimental analysis (for both datasets), it is observed that, the designed system with all three features sets give 98 to 100 % correct identification rate under *matched* condition (same sensors/not shown in the results). Fig.4 shows the evaluation results using Hindi dataset, whereas Fig.5 illustrates the results obtained on IITG database. From Fig. 4 (a) it is observed that all three features in its baseline form are very much sensitive to sensor mismatch. Adding Gaussian non-linearity and appending dynamic information to spectral improves the percentage identification rate appreciably as observed from Fig.4 (b). Inclusion of spectral subtraction in combination with feature normalization further improves the accuracy of the system using to 100 % in the case of proposed (NDSF) features.



(a) (b)

Fig.4 Speaker Identification Performance (a) without and (b) with dynamic features (Hindi Dataset)

Under matched and clean training and testing condition (Headset-Headset), MFCC features found to give slightly better results (1 % to 2 %) than LPCC and NDSF features. Fig. 5 (a) and (b) compares the performance of various features for *sensor mismatch* condition using IITG-MVSR database. As observed from Fig.5 (a), the acoustic mismatch caused due to mismatch in devices used for training and testing, drop down the recognition rate drastically, especially for the case of MFCCs and LPCCs. Also it is apparent from plots that NDSF features in its baseline form (without use of CMN and SS) gives better percentage identification for all four cases of mismatch than that of MFCC and LPCC features (Fig.5.a). Addition of temporal (dynamic) information to the three feature sets(Fig. 5.b), resulted in significant improvement in all four cases of mismatch. Dynamic spectral features are found to perform much better especially for the case of Headset-Digital voice recorder. Inclusion of spectral subtraction in combination with cepstral mean normalization to NDSF features further improves the accuracy to almost 100 %.

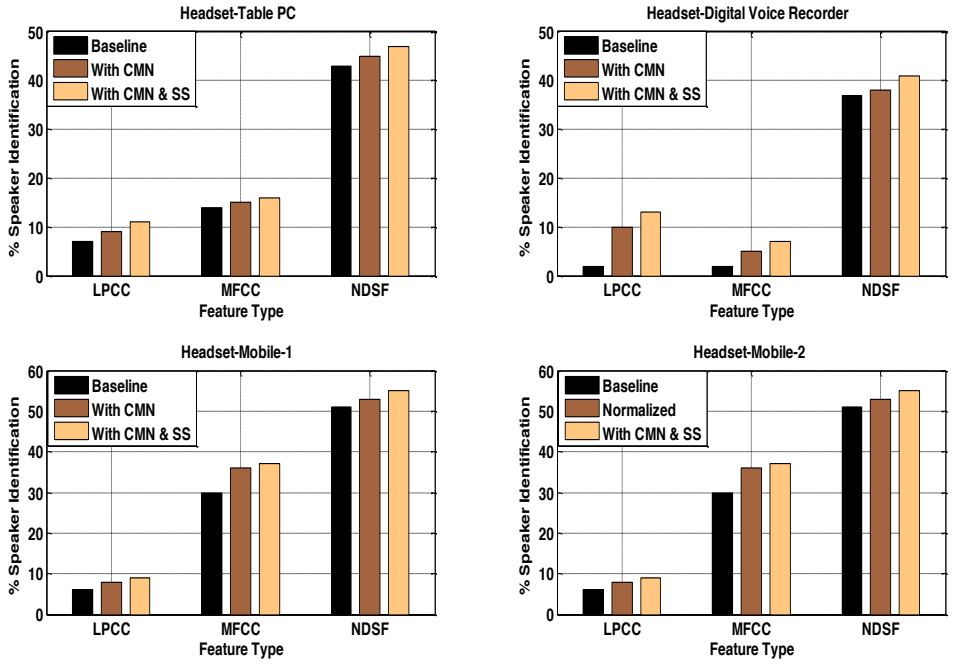


Fig.5 (a) Speaker Identification Performance of features without dynamic features (IITG Dataset)

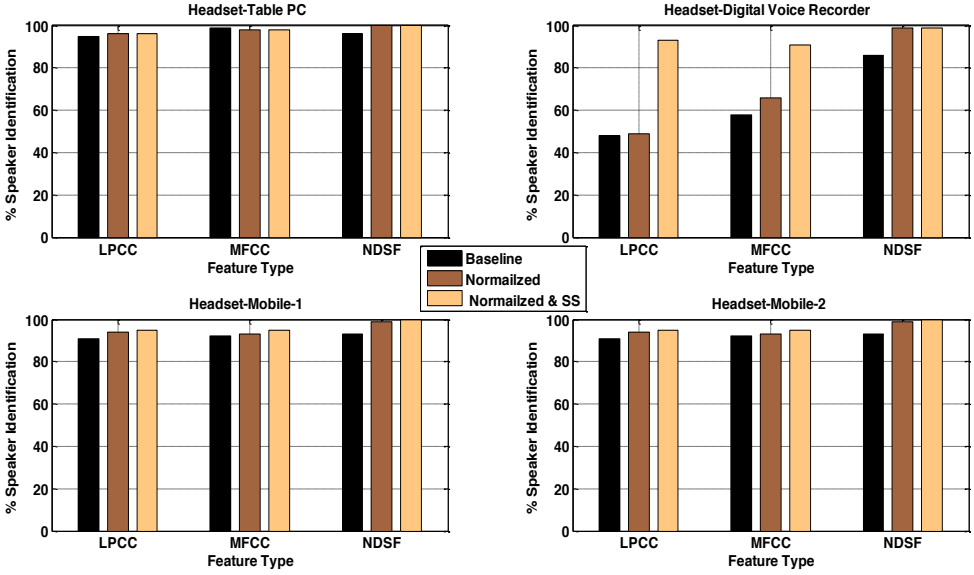


Fig.5 (b) Speaker Identification Performance of features with dynamic features (IITG Dataset)

6. Conclusion

In this work, a new set of features set called Normalized Dynamic Spectral Features (NDSF) is proposed and investigated for sensor mismatch condition. Use of dynamic spectral information and Gaussianized feature warping showed greater contribution towards the robustness in mismatch condition. The proposed feature set is observed to be more robust than cepstral features like MFCCs and LPCCs. The case of sensor mismatch is investigated in this work and gives the promising results. Further work will be to examine the performance of NDSF features in various other real world mismatch conditions.

Acknowledgement

The authors would like to thank EMST Lab IIT Guwahati and Department of Information Technology, Government of India and TIFR, Mumbai for providing speech database for the research work. Authors would also like to thank Department of Technology, Shivaji University, Kolhapur for their kind support in doing the research work.

References

1. Richard J. Mammone ,Xiaoyu Zhang ,Ravi P. Ramachandran: Robust Speaker Recognition: A Feature based approach, *IEEE Signal Processing Magazine,* September 1996.
2. J. P. Campbell, Jr.: Speaker Recognition: A Tutorial, Proceedings of The IEEE, Vol.85, No.9, pp.1437-1462, Sept.1997.
3. Tomi Kinnunen, Haizhou Li: An overview of text independent speaker recognition, from features to supervectors, *Speech Communication,* July 2009.
4. Marcos Faundez-Zanuy and Enric Monte-Moreno: State-of-the-art in Speaker Recognition *, IEEE A&E Systems Magazine*, May 2005
5. S. Furui: Speaker-independent isolated word recognition based on emphasized spectral dynamics, *Proc. ICASSP*, 1986.
6. Kshitiz Kumar, Chanwoo Kim and Richard M. Stern: Delta spectral cepstral coefficients for robust speech recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2011): 4784-787.
7. Douglas A Raynolds: Automatic speaker recognition using Gaussian mixture speaker models, The *LINCOLN Laboratory Journal*, Vol.8 No.2 , pp.173-192,1995.
8. Ji Ming, Timothi J Hazen, James R Glass, Douglas A Raynolds :Robust speaker recognition in noisy conditions*, IEEE Transaction on Audio, Speech and Language Processing*, Vol.15, No.5, July 2007.
9. K. Sri Rama Murty and B. Yegnanarayana: Combining evidence from residual phase and MFCC features for speaker recognition , *IEEE Signal Processing Letters*, Vol.13, No.1, January 2006.
10. Andre G Adami, Radu Mihaescu, Douglas A Raynolds, John j Godfrey: Modeling prosodic dynamics for speaker recognition, *ICASSP*-2003.
11. Robert Tongeri and Daniel Pullela,: An overview of speaker identification: Accuracy and robustness issues, *IEEE Circuits and Systems Magazine*, 2011, pp.1531-1569, Second quarter 2011.
12. Danoush Hosseinzadeh and Sridhar Krishnan: On the use of complementary features for speaker recognition*, Eurasip Journal of Advances in Signal Processing*, 2008.
13. Pervouchine V., Leedham G., Zhong H., Cho D. and Li H: Comparative study of several novel acoustic features for speaker recognition, Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing, pp. 220-223,2008.
14. Seyed Omid Sadjadi and John H.L. Hansen: Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions, ICASSP 2011.
15. Sharada V Chougule and Mahesh S Chavan: Channel Robust MFCCs for Continuous Speech Speaker Recognition, *Advances in Signal Processing and Intelligent Recognition Systems,* Spinger International Pub.2014,Volume 264, 2014, pp 557-568
16. Sriram Ganapathy*,* Sri Harish Mallidi*,* and Hynek Hermansky: Robust Feature Extraction Using Modulation Filtering of Autoregressive Models, IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol 22, No. 8, August 2014, pp.1285-1295
17. Biagetti G.,Crippa P.,Curzi A.,Orcioni S.andTurchetti C.: Speaker identification with short sequence of speech frames, Proceedings of International Conference on Pattern Recognition Applications and Methods (ICPRAM-2015), pp.178-185.
18. D O'Shaughnessy, Speech Communication: Human and Machine, Reading, MA: Addison-Wesley, 1987.
19. Saeed V. Vaseghi ,"Advanced Digital Signal Processing and Noise Reduction ", Second Edition, *John Wiley & Sons* Ltd.2000.
20. Tomas F. Quatieri : Discrete Time Speech Signal Processing, Principles and Practice*, Pearson Education* 2006.
21. http://mathworld.wolfram.com/NormalDistribution.html
22. Y. Linde, A Buzo and R M Gray: :An algorithm for feature vector quantizer design", *IEEE transaction on Communication*, vol.28, No.1, pp.84-95,1980
23. http: //www.iitg.ernet,in/ece/emstlab
24. Samudravijaya K, P.V.S.Rao, and S.S.Agrawal: Hindi Speech Database, Proceedings of *International Conference on Spoken Language Processing,* 2000, China.